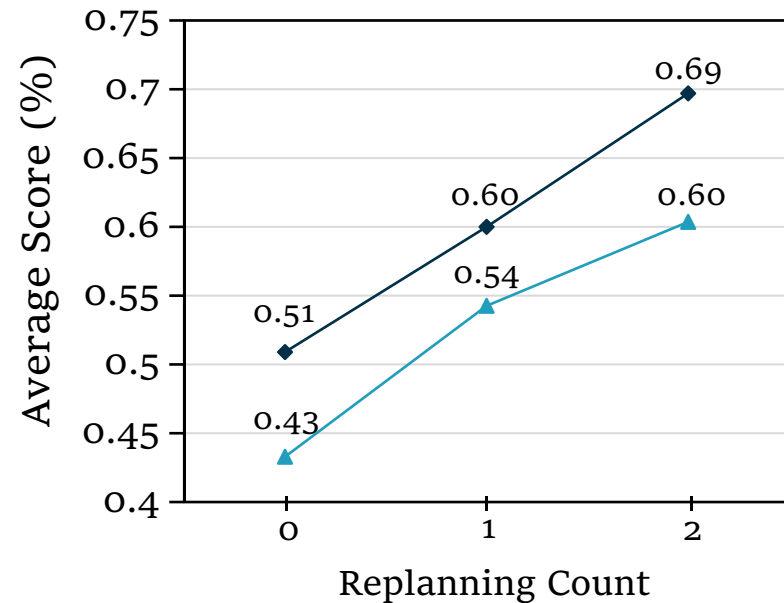


(a) Performance by Capability



(b) Test-time scaling